On the Statistical Performance of Hierarchical Bayes MNL Conjoint Models: Findings from Simulation Studies

DISSERTATION

zur Erlangung des Doktorgrades
der Wirtschaftswissenschaften

vorgelegt von

Maren Hein
aus Celle

genehmigt von der
Fakultät für Energie- und Wirtschaftswissenschaften
der Technischen Universität Clausthal

Tag der mündlichen Prüfung:

12.04.2017

Vorsitzende der Promotionskommission: Prof. Dr. rer. pol. Inge Wulf

Hauptberichterstatter: Prof. Dr. rer. pol. Winfried J. Steiner

Berichterstatter: Prof. Dr. rer. pol. Bernhard Baumgartner

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Lehrstuhl für Betriebswirtschaftslehre und Marketing des Instituts für Wirtschaftswissenschaft der Technischen Universität Clausthal. Im Januar 2017 wurde die Arbeit von der Fakultät für Energie- und Wirtschaftswissenschaften als Dissertation angenommen. Viele Menschen haben mich auf diesem Weg begleitet und auf verschiedene Art und Weise zum Entstehen der Dissertation beigetragen, wofür ich mich ganz herzlich bedanken möchte.

Zuallererst möchte mich bei meinem Doktorvater Prof. Dr. Winfried Steiner für die ausgezeichnete Betreuung bedanken. Vielen Dank für die wertvolle fachliche Unterstützung, das große Vertrauen in meine wissenschaftliche Arbeit und die unglaubliche Motivationsgabe, die mich immer wieder angetrieben und maßgeblich zum Gelingen dieser Arbeit beigetragen hat.

Danken möchte ich auch Prof. Dr. Bernhard Baumgartner von der Universität Osnabrück für sein Interesse an meiner Arbeit und die Übernahme des Zweitgutachtens.

Darüber hinaus möchte ich Peter Kurz von TNS Infratest München meinen Dank aussprechen. Peter Kurz war mir bei Fragen und Problemen aufgrund seiner umfangreichen Kenntnisse und Erfahrungswerte auf dem Gebiet der Conjointanalyse eine große Hilfe. Ohne ihn wäre eine enge Verbindung von Theorie und Praxis nicht möglich gewesen. Vielen Dank für die angenehme, konstruktive und bereichernde Zusammenarbeit.

Mein Dank gilt weiterhin den Mitarbeitern des Instituts für Wirtschaftswissenschaft und insbesondere meinen Kolleginnen und Kollegen der Abteilung für Betriebswirtschaftslehre und Marketing. Zusammen mit ihnen habe ich nicht nur die alltäglichen Herausforderungen des Lehrstuhlalltags mit Freude gemeistert, sondern konnte mir stets gewiss sein, auch bei Problemen ein offenes Ohr und Zeit für Gespräche zu finden. Neben der fachlichen Zusammenarbeit war es daher auch in persönlicher Hinsicht eine unvergleichliche Zeit, die ich nicht missen möchte.

Ein ganz besonderer Dank gilt meinen Eltern, die immer für mich da sind und ohne deren Begleitung meines Lebens- und Bildungsweges, diese Arbeit vermutlich niemals hätte geschrieben werden können. Schließlich und keineswegs zuletzt danke ich meinem Freund Benjamin, der mich darin bestärkt hat, das Projekt Promotion anzufangen, auch wenn damit

weitere Jahre der räumlichen Trennung verbunden waren. Vielen Dank für deine Geduld, die vielen aufmunternden Worte während der abendlichen Telefonate und deine uneingeschränkte Unterstützung!

Clausthal-Zellerfeld, Januar 2017 Maren Hein

Contents

List of tables and figures

List of abbreviations

ANOVA	analysis of variance
BTL	Bradley-Terry-Luce
CBC	choice-based conjoint
cf.	confer
d.f.	degrees of freedom
e.g.	exempli gratia (Latin) = for instance
et al.	et alia (Latin) = and others
excl.	exclusive
FC	first choice (rule)
FM	finite mixture
HB	hierarchical Bayes
i.e.	id est (Latin) = that is
IIA	independence of irrelevant alternatives
LC	logit choice (rule)
MAE	mean absolute error
MAPE	mean absolute percentage error
MCMC	Markov Chain Monte Carlo
MNL	multinomial logit model
MPE	mean percentage error
MSE	mean squared error
p.	page
RAE	relative percentage error
RFC	randomized first choice (rule)
RMSE	root mean square error
vs.	versus
w.r.t.	with respect to
%TrueBetas	percentage of true part-worths that are covered by 95% Bayesian credible intervals

Chapter 1
Introduction

The primary aim of most companies is profit maximization and growth. However, due to an increasing economic globalization companies have to be able to cope with a growing competitive pressure. In order to position themselves effectively on the market, companies are forced to create products able to satisfy consumer needs and preferences, better than competition. The inability to adequately satisfy consumer preferences and fulfill consumer needs is one of the main reasons for the failure of new or modified products. A widely applied method for measuring, analyzing and predicting consumer preferences is conjoint analysis. Since its introduction to marketing research (Green and Rao 1971; Green and Srinivasan 1978) conjoint analysis has become a popular method in marketing science and practice. Based on the pioneering work by McFadden (1974) on discrete choice models and Louviere and Woodworth (1983) who integrated conjoint and discrete choice approaches, especially Choice-Based Conjoint (CBC) analysis has evolved as the most widely used conjoint method in marketing theory and practice. CBC analysis combines "the best features of discrete choice models and conjoint analysis" (Cohen 1997, p.14). In general, conjoint analysis is a data collection technique using experimental designs. Based on random utility theory proposed by Thurstone (1927) discrete choice models are statistical methods for analyzing choice responses and can be applied for example to CBC data (Cohen 1997).[1] The objective of CBC analysis is to estimate consumers' preference structures (part-worth utilities) by asking individuals to choose among different sets of alternatives (choice tasks), where each alternative is described by several attributes and attribute levels. Thus, in contrast to traditional conjoint analysis where individuals are asked to rate or rank a single set of alternatives, CBC analysis closely mimics what individuals do in real environments because when purchasing products they make choices and probably do not rank or rate product alternatives (Louviere 1988).[2] A serious limitation in early applications of CBC analysis was that part-worths could not be estimated at the individual, but only at the aggregate or segment level, because choice data provides less information than rankings- or ratings-based conjoint data (Huber and Train 2001; Green, Krieger, and Wind 2001). The revolution occurred with the availability of Hierarchical Bayes (HB) estima-

[1] For a detailed distinction between discrete choice models and conjoint analysis see Louviere, Flynn, and Carson 2010.
[2] For an overview of advantages of CBC analysis over traditional conjoint analysis see for example Cohen 1997.

tion procedures (e.g., Allenby, Arora, and Ginter 1995; Allenby and Ginter 1995; Lenk et al. 1996). The application of HB methods nowadays allows the estimation of reliable individual-level part-worths so that it is possible to recover respondents' heterogeneous preferences from choice data (Allenby et al. 2005). In particular, HB combines information from two sources to estimate individual-level part-worths: i) each individual's choice data represented by a model of decision making and ii) a model that describes the distribution of preferences across all respondents (Allenby and Rossi 2006).

In the present work we focus on CBC analysis using HB for the estimation of part-worth utilities. Specifically, we conduct three extensive simulation studies in order to explore the statistical performance of HB-CBC models.

The objective of the first simulation study is to examine if there exists a limit for parameter settings in CBC studies. With the use of HB it is possible to estimate individual part-worths from a technical point of view even when there are more parameters than observations. As a consequence market researchers are confronted with the problem that clients desire to include more and more attributes while keeping the choice task manageable. Therefore, the question is how many attributes, how few respondents, or how few choice tasks per respondent can be considered in a CBC model in order to ensure still good estimation and prediction results.

The second simulation study focuses on the prediction of preference shares. Since from a managerial point of view part-worths resulting from CBC analysis are rather abstract and difficult to interpret shares of preference are derived subsequently that are easy to understand and much more appealing. In order to predict which products respondents would choose in a (hypothetical) market scenario, different choice rules can be used that relate respondents' utilities to expected individual choice probabilities. Those choice probabilities can be aggregated across respondents to obtain the share of respondents who prefer one product compared to the other competing items. However, each choice rule has its pros and cons. As a consequence choice share predictions can be different depending on the applied choice rule and may lead to wrong managerial decisions. Thus, the second study wants to shed more light on the question which choice rule should be used in order to predict preference shares as accurate as possible. In particular, the special focus here lies on the use of HB draws combined with first choice simulations for preference share predictions.

As mentioned above, the key strength of HB is its ability to provide individual part-worth estimates given only relatively few observations per respondent. In order to be able to stably

estimate individual part-worths, HB inference requires prior beliefs about the unknown part-worths. Therefore, in our third simulation study the objective is to investigate the impact of these prior beliefs on the performance of HB-CBC models. To be precise the goal of the study is to substantially contribute to the question how HB prior parameter settings (i.e. the prior variance and the prior degrees of freedom) affect the performance of HB-CBC models.

An overview of the most important studies in the context of these three particular fields of research addressed in this thesis is given in Table 1.1. In previous simulation studies related to conjoint analysis researchers used synthetic data to compare the performance of different conjoint segmentation methods (Vriens, Wedel, and Wilms 1996) or of conjoint models estimated at different levels of aggregation (Andrews, Ansari, and Currim 2002; Backhaus, Hillig, and Wilken 2007; Chakraborty et al. 2002). Here, we use simulated data in order to be able to explore the statistical performance of HB-CBC under systematically varying conditions. The great advantage of synthetic data is that the true parameters are known and can be compared to the estimated ones. Thus, the performance of HB-CBC can be assessed. With regard to the performance of HB estimation Backhaus, Hillig, and Wilken (2007) found that the HB approach performs best across competing choice-based conjoint models (traditional CBC, latent class CBC, HB-CBC) and ratings-based limit conjoint models (limit conjoint, limit FM conjoint, limit HB conjoint models) independently from the data collection process (i.e. choice or rating data). As well, the simulation study by Wirth (2010) revealed that HB-CBC works very well even under sparse data conditions. The present work will show that for simple CBC settings HB estimation proves to be quite robust, but when the CBC design is already complex HB is starting to collapse under certain conditions.

With regard to choice rules comparisons previous studies focused on the traditional First Choice (FC) rule and probabilistic choice rules like the Logit Choice (LC) rule or the Bradley-Terry-Luce (BTL) rule (e.g. Finkbeiner 1988; Green and Krieger 1988; Elrod and Kumar 1989), or new (modified) approaches (compare Baier and Gaul 2007; Tsafarakis, Grigoroudis, and Matsatsinis 2011). Further, although the use of HB estimation techniques for CBC data has been established since the mid 90’s and also the Randomized First Choice (RFC) rule was introduced to correct for product similarity, only few studies analyzed the performance of using HB draws or the RFC rule for choice share predictions (Huber, Orme, and Miller 1999; Orme and Baker 2000; Baier and Polasek 2003; Arenoe 2003). In particular, only two studies provide findings on the predictive performance of using HB draws for preference simulation and both studies used empirical data. Baier and Polasek (2003) presented an empirical study

for metric conjoint data. They assessed the predictive accuracy of simulations from HB random draws compared to traditional procedures for preference simulations. The authors found that HB draws are superior as compared to probabilistic choice rules. The study conducted by Orme and Baker (2000) is the only one that compared predictions based on HB draws to predictions based on the RFC rule. RFC simulations turned out to be slightly more accurate for the data sets considered than those using HB draws. In the present thesis we consider traditional choice rules as well as the RFC rule and HB draws for simulating shares of preference among competitive product concepts. To the best of our knowledge there is no study based on synthetic choice-based conjoint data that systematically explores the accuracy of preference share predictions by HB draws in comparison to traditional choice rules like the First Choice rule, the Logit Choice rule, as well as to the RFC rule. Our results clearly show the superiority of HB draws combined with first choice simulations that lead to the lowest prediction errors across all choice rules considered.

Related studies concerning HB prior settings mainly concentrated on priors in the context of sparse CBC data sets (Pinnell and Fridley 2001; Orme 2003; McCullough 2009; Lenk and Orme 2009). Orme (2003) found that default settings for the HB covariance matrix priors lead to overfitting for sparse data sets and pointed out that results improved by adjusting priors to be more informative. In addition, McCullough (2009) showed that default prior settings are not optimal for sparse CBC data sets. Furthermore, evidence that more informative priors can improve the estimation for sparse data sets is provided by Lenk and Orme (2009). A meta-analysis of 50 commercial CBC data sets conducted by Orme and Williams (2016) further revealed that the optimal prior settings depend on the data set characteristics. Across CBC data sets the results showed that the optimal prior variance setting ranged from 0.1 to 1.6. Hence, they proposed a default value of one. In the present study we find that overfitting with respect to parameter recovery and model fit becomes evident for a prior variance of 4. The main finding of our simulation study is that the predictive performance of HB turns out to be only slightly affected by the prior variance even in the most extreme cases, i.e. if we go far beyond a prior variance of 4. Further, results show that the prior degrees of freedom settings play only a negligible role not having any noticeable impact on the performance of HB.

Table 1.1: An overview of the most important studies related to the fields of research in this thesis

Paper	Conjoint Approach	Data Basis	Objectives	Main Findings	Field of Research in this Thesis
Andrews, Ainslie, Currim (2002)	Choice-based (scanner panel data, no conjoint approach)	Synthetic panel data	Performance comparison of HB mixed logit choice models and Finite Mixture logit choice (FM) models	HB has an advantage with regard to model fit, but performs poorly under certain extreme conditions	Study 1: Limits for parameter settings in CBC analysis
Andrews, Ansari, Currim (2002)	Metric	Synthetic conjoint data	Performance comparison of HB conjoint and FM conjoint models	HB and FM perform equally well with regard to part-worth recovery and predictive accuracy; HB better than FM in terms of model fit	Study 1: Limits for parameter settings in CBC analysis
Arenoe (2003)	Choice-based	Empirical conjoint data	Analysis of the effects of using different estimation and prediction techniques on the external validity of CBC (choice rules: FC, RFC)	RFC with both product and attribute variability and RFC with only product variability better than FC	Study 2: Using HB draws for improving predictions in conjoint simulations
Backhaus, Hillig, Wilken (2007)	Metric as well as choice-based	Synthetic conjoint data	Performance comparison of different variants of choice-based conjoint models and ratings-based limit conjoint models	Independently from data collection process HB proves to be robust and shows best results	Study 1: Limits for parameter settings in CBC analysis

Paper	Conjoint Approach	Data Basis	Objectives	Main Findings	Field of Research in this Thesis
Baier, Gaul (2007)	Metric (graded paired comparisons)	Synthetic as well as empirical conjoint data	Introduction of a new approach based on an ideal vector model and analysis of its predictive performance in comparison to traditional approaches	New approach better than traditional choice rules	Study 2: Using HB draws for improving predictions in conjoint simulations
Baier, Polasek (2003)	Metric	Empirical conjoint data	Comparison of HB draws to traditional procedures for preference share simulations (FC, BTL, LC)	Predictive performance of HB draws better than based on BTL and LC, and slightly worse than under FC	Study 2: Using HB draws for improving predictions in conjoint simulations
Chakraborty, Ball, Gaeth, Jun (2002)	Metric as well as choice-based	Synthetic conjoint data	Performance comparison of individual-level ratings-based and aggregate-level choice-based models	Performance of the methods depends on underlying conditions (level of consumer heterogeneity, product similarity, type of consumer choice rule applied, amount of error in utilities)	Study 1: Limits for parameter settings in CBC analysis
Huber, Orme, Miller (1999) (similar Orme and Huber 2000)	Choice-based	Empirical conjoint data	Examination of the ability of the RFC choice rule under different levels of variability to correctly reflect similarity effects in market simulations	Adding attribute variability is crucial for a good performance of RFC	Study 2: Using HB draws for improving predictions in conjoint simulations

Paper	Conjoint Approach	Data Basis	Objectives	Main Findings	Field of Research in this Thesis
Lenk, Orme (2009)	Choice-based	Synthetic conjoint data	Analysis of the effect of priors for variances and covariance matrices in the context of sparse data sets	Concerning CBC studies with categorical attributes and effects-coding, results are variant to the choice of the omitted level when using a default prior covariance matrix; estimation of part-worths under effects coding is not biased when using the proposed effects-coding prior	Study 3: Analyzing HB covariance matrix prior settings
McCullough (2009)	Choice-based	Empirical conjoint data	Comparison of default and advanced forms of latent class CBC and HB-CBC applied to sparse data sets; with regard to HB default prior settings and adjusted priors are compared	Latent class and HB models perform similar well in default and more advanced forms; for sparse data sets the performance of HB improves in terms of holdout hit rates when adjusted prior settings are used	Study 3: Analyzing HB covariance matrix prior settings
Orme, Baker (2000)	Choice-based	Empirical conjoint data	Comparison of the predictive accuracy of preference share simulations based on HB draws vs. RFC choice rule	RFC with attribute variability slightly more accurate than HB draws	Study 2: Using HB draws for improving predictions in conjoint simulations
Orme (2003)	Choice-based	Empirical conjoint data	Illustration that default prior settings have to be changed in certain circumstances	Under extreme conditions (many parameters, sparse data) default prior setting are suboptimal leading to overfitting; performance of HB improves by adjusting priors	Study 3: Analyzing HB covariance matrix prior settings

Paper	Conjoint Approach	Data Basis	Objectives	Main Findings	Field of Research in this Thesis
Orme and Williams (2016)	Choice-based	Empirical conjoint data	Meta-analysis of 50 commercial CBC data sets to find optimal HB prior settings	Optimal prior settings depend on the data set characteristics; optimal prior variance ranges from 0.1 to 1.6 so that a default value of 1 should be used; general tendency that as the number of attributes in CBC studies increases, the optimal prior variance tends to decrease and vice versa	Study 3: Analyzing HB covariance matrix prior settings
Tsafarakis, Grigoroudis, Matsatsinis (2011)	Metric	Synthetic as well as empirical conjoint data	Introduction of a new modified BTL approach and comparison of its performance to traditional choice rules	New modified BTL approach (tuning approach) only slightly better than generalized BTL and better than traditional choice rules	Study 2: Using HB draws for improving predictions in conjoint simulations
Vriens, Wedel, Wilms (1996)	Metric	Synthetic conjoint data	Performance analysis of different one-stage and two-stage conjoint segmentation methods	Two-stage clustering procedures are outperformed by integrated methods; preferable method depends on primary purpose of the study	Study 1: Limits for parameter settings in CBC analysis
Wirth (2010)	Choice-based	Synthetic conjoint data	Evaluation of the performance of the HB-Best-Worst-CBC approach as new conjoint variant in comparison to the standard HB-CBC and a non-HB approach	Standard HB-CBC performs well with regard to part-worth recovery and prediction of individual choice behavior; HB-Best-Worst-CBC outperforms HB-CBC with respect to prediction of preference shares	Study 1: Limits for parameter settings in CBC analysis

Objectives and outline of this thesis

In the first part (simulation study 1), the main focus lies on the detailed investigation as to when HB estimation in CBC analysis reaches its limits. In order to examine if there exists a limit for parameter settings in CBC studies, we design a simulation study using synthetic choice-based conjoint data.[3] That way, we are able to explore how few respondents, how few choice tasks per respondent, or how many attributes one can consider in a HB-CBC model before the statistical model performance gets considerably worse. The simulation design (choice of experimental factors and data generation process) except for some modifications closely follows the study designs as proposed by Vriens, Wedel, and Wilms (1996), Andrews, Ainslie, and Currim (2002), Andrews, Ansari, and Currim (2002), and Wirth (2010) (compare Table 1.1). Accordingly, the statistical performance of HB is evaluated under experimentally varying conditions based on seven experimental factors (among them the number of attributes, choice tasks, and respondents, as well as the number of attribute levels, number of alternatives per choice task, the sample structure, and the amount of error) using criteria for goodness-of-fit, parameter recovery and predictive accuracy. In addition, analyses of variance (ANOVAs) are conducted to assess the impact of the experimental factors on the measures of performance. We further perform a sensitivity study with still more extreme level settings for some of the experimental factors.

In the second part (simulation study 2), we focus on the application of CBC analysis to conduct market simulations. The application of HB estimation in CBC analysis offers the opportunity to directly use HB random draws for those market simulations. However, only few studies have been conducted to assess the predictive accuracy of simulations from HB draws (compare Table 1.1). In this context (using the HB-CBC model to estimate individual part-worth utilities) we systematically compare market share predictions based on the following choice rules: (1) the First Choice rule, (2) the Logit Choice rule, (3) the Randomized First Choice rule, and (4) HB random draws combined with first choice simulations. Here, based on the data generation process used in simulation study 1, we also design a simulation study in order to examine the conditions under which one of these choice rules recovers preference shares better than the other. Further, we investigate the power of the different choice rules to handle predictions for similar alternatives and therefore assess how well they tolerate the Ir-

[3] In all three studies presented in this thesis individual-level part-worths are generated and respondents' choices are simulated by using the software R 2.15.2 and higher versions (R Core Team 2012). Based on the simulated respondents' choices the individual-level part-worths (as well as the population means and the covariance matrix) are then re-estimated using HB as implemented in the Sawtooth Software.

relevance of Independent Alternatives (IIA) property (cf. Orme and Huber 2000; Orme and Baker 2000). To investigate how well the choice rules account for the IIA property, two different holdout choice scenarios containing three alternatives each are carefully designed under each treatment. In order to assess the potential IIA bias one holdout task is designed to have two of the three alternatives extremely similar to each other. The second holdout task consists of alternatives that are all different from each other. Analogous to simulation study 1, the performance of the four choice rules is evaluated under experimentally varying conditions based on four experimental factors (the choice rule, number of choice tasks, sample structure as well as the model complexity) using statistical criteria for predictive accuracy. Analyses of variance are conducted to assess the impact of the experimental factors on the measures of predictive accuracy.

So far, the simulation studies are conducted by using default software settings for HB estimation. Specifically, HB prior parameter settings (prior variance and prior degrees of freedom) have not been changed from the default values. However, it may depend on the characteristics of the respective data set whether the default HB prior parameter settings are appropriate (compare Table 1.1). Therefore, the third part (simulation study 3) goes beyond the standard HB prior settings and presents a simulation study, also based on the data generation process of simulation study 1, to contribute to the question how HB prior parameter settings affect the performance of HB-CBC models. We investigate the influence of the HB priors by systematically varying further experimental factors, such as the number of respondents, the number of choice tasks and the number of parameters to be estimated. Overall, the statistical performance of HB is evaluated under experimentally varying conditions based on seven experimental factors using criteria for goodness-of-fit, parameter recovery and predictive accuracy. Moreover, analyses of variance are conducted to assess the impact of the experimental factors on the measures of performance. In addition, a sensitivity analysis is performed to test even higher prior variance levels.

As this is a short version of the thesis, we briefly introduce the HB model for choice-based conjoint analysis used in all three simulation studies in Chapter 2. Next, we describe the design of the three simulation studies in Chapter 3. Subsequently, the findings from the simulation studies are summarized in Chapter 4. Finally, the thesis closes with a discussion of the limitations and with an outlook on future research perspectives.

For a detailed description of the design of the simulation studies as well as the results of each simulation study the reader can contact the corresponding author for an extended version of this thesis.

Chapter 2
The HB model for choice-based conjoint analysis

The HB method can be classified as a random-effects model in which the parameters are assumed to vary across respondents according to a probability distribution. In classical random-effects models the parameters of the probability distribution can be estimated. However, it is not possible to draw inferences about individual-level parameters (e.g., Rossi and Allenby 1993; Allenby and Ginter 1995; Rossi, Allenby, and McCulloch 2005). In contrast, the HB random-effects model is characterized by a hierarchical structure that allows the estimation of individual-level parameters using both information from the probability distribution of all individuals and each individual´s choice data.

At the individual level (lower level) of the hierarchical structure the probability of each individual´s choice is modeled by a multinomial logit model. The MNL model is based on the assumption that error terms are independently identically Gumbel-distributed[4] with location parameter η equal to zero and scale factor μ. The variance of the error term can be manipulated via the scale factor. The standard error variance of the MNL corresponds to a scale factor of *one* $(\mu = 1)$, to double the error variance the scale factor is set to the square root of two $(\mu = \sqrt{2})$.[5] The higher the variance the higher the stochastic component and the worse should the choice behavior be representable. The MNL model can be expressed as:

$$P_n(j') = \frac{\exp(\frac{1}{\mu}\beta_n x_{j'})}{\sum_{j=1}^{J} \exp(\frac{1}{\mu}\beta_n x_j)},$$

where $P_n(j')$ is the probability that the *n*-th respondent chooses the *j'-th* alternative in a particular choice task, β_n represents the vector of part-worths for the *n*-th respondent, $x_{j'}$ is a

[4] The Gumbel distribution is also known as the extreme value distribution of type I:
$G(x) = \exp(-\exp(-\frac{1}{\mu}(x-\eta)))$, $\mu > 0$.

[5] The variance of the Gumbel distribution is $\frac{\pi^2\mu^2}{6}$, which amounts to 1.645 if μ =1, and to 3.290 if $\mu = \sqrt{2}$.

dummy vector for the attribute levels of alternative j', and $\mu > 0$ represents the scale parameter of the logit model.

At the population level (upper level) the Bayesian normal model is given by the first-stage prior

$$\beta_n \sim \mathrm{N}(\overline{\beta}, V_\beta)$$

and the second-stage priors

$$\overline{\beta} \sim \mathrm{N}(b_0, S_0) \quad \text{and} \quad V_\beta \sim \mathrm{IW}(\upsilon, \Lambda).$$

In the basic hierarchical normal model the multivariate normal distribution is typically used as first-stage prior where $\overline{\beta}$ represents the vector of means of the distribution of the individuals' part-worths, and V_β is the covariance matrix that captures the extent of heterogeneity (as well as the correlation in the part-worths) across individuals (Train 2003; Rossi and Allenby 2003; Rossi, Allenby, and McCulloch 2005).

To estimate the parameters $\overline{\beta}$ and V_β of the first-stage prior, hyperprior distributions (second-stage priors) have to be specified. It is common to assume that the prior on $\overline{\beta}$ is represented by a normal distribution with mean b_0 and variance S_0, and that the prior on the covariance matrix V_β is inverse Wishart distributed with υ degrees of freedom and a covariance (scale) matrix Λ (Train 2003; Rossi, Allenby, and McCulloch 2005). The second-stage priors are usually set to be very diffuse (i.e. representing little information) to let $\overline{\beta}$ and V_β be determined primarily by the data (Rossi, Allenby, and McCulloch 2005; Train 2003). The normal prior on $\overline{\beta}$ becomes more spread out and flat by raising the variance S_0 of the prior, and the inverse Wishart prior on V_β becomes more diffuse with lower υ and larger elements of the covariance matrix Λ (Rossi, Allenby, and McCulloch 2005). β_n, $\overline{\beta}$ and V_β are unknown and have to be estimated using the information from the underlying choice data. Thus, the prior information is combined with the observed choice data to estimate the posterior distributions (Bayes theorem). To sample from the complex joint posterior distribution of the unknown parameters β_n, $\overline{\beta}$ and V_β conditioned on the observed data Markov Chain Monte Carlo (MCMC) simulation methods like the Metropolis-Hastings (MH) algorithm and the

Gibbs sampler are used. In particular, instead of taking draws from the joint posterior for all parameters simultaneously a sequence of draws is generated by an iterative process. That means draws are taken from the posterior for one parameter at a time conditional on values of the other parameters (Train 2003). After a burn-in phase where the Markov chain has converged, the draws are usually averaged to calculate point estimates of the parameters. Alternatively, individual HB draws can be used for subsequent steps (e.g., preference simulations).

Chapter 3
Design of the simulation studies

3.1. Data

In all three simulation studies we experimentally manipulated various factors and each factor was varied at several levels. The experimental factors are chosen following previous simulation studies and the synthetic data generation is also similar to that of previous simulation studies (Vriens, Wedel, and Wilms 1996; Andrews, Ainslie, and Currim 2002; Andrews, Ansari, and Currim 2002; Wirth 2010). An overview of the relevant experimental factors and factor levels related to the three particular simulation studies conducted in this thesis is given in Table 3.1.

The synthetic data are generated according to the HB approach. The standard HB approach assumes that (a) at the individual respondent level choices are modeled via the MNL model and that (b) at the population level individual-level part-worth vectors β_n follow a multivariate normal distribution $\beta_n \sim N(\overline{\beta}, V_\beta)$ with population mean beta vector $\overline{\beta}$ and covariance matrix V_β. Accordingly, we randomly generated the true part-worths for each respondent as to follow a multivariate normal distribution. Note that the length of a part-worth vector β_n and consequently the length of the population mean beta vector $\overline{\beta}$ depends on factors 1 and 2 (number of attributes, number of attribute levels) in study 1 respectively on factor 9 (simple or complex scenario) in study 2 and varies between 12 and 48 part-worth parameters. 'Simple scenario' means that relative few parameters need to be estimated (6 attributes, 3 attribute levels), whereas the 'complex scenario' is characterized by a relative high number of parameters to be estimated (12 attributes, 5 attribute levels). In simulation study 3 the length of the individual-level part-worth vector β_n and the population mean beta vector $\overline{\beta}$ depends only on factor 1 (number of attributes) and varies between 24 and 56 part-worth parameters as the number of attribute levels is held constant at a value of 5 here.[6]

[6] Note that for H attributes with I levels each, H times (I-1) part-worths need to be estimated independent whether a dummy- or effects-coding is used.

Table 3.1: Overview of experimental factors and factor levels

Factor		# Factor levels	Factor levels	Related simulation study
1.	Number of attributes	4	6, 8, 10, 12	Study 1
		3	6, 10, 14	Study 3
2.	Number of attribute levels	3	3, 4, 5	Study 1
3.	Number of choice tasks (excl. holdout tasks)	3	11, 13, 15	Study 1
		3	7, 11, 15	Study 2
		2	7, 11	Study 3
4.	Number of respondents	3	500, 1000, 1500	Study 1
		3	100, 500, 1000	Study 3
5.	Sample structure	2	homogeneous, heterogeneous	Study 1, Study 2, Study 3
6.	Error variance	2	standard (1.645), high (3.290)	Study 1, Study 3
7.	Number of alternatives per choice task (excl. none option)	3	3, 4, 5	Study 1
8.	Choice rule	4	FC, LC, RFC, HB draws + FC	Study 2
9.	Model complexity	2	simple, complex	Study 2
10.	Prior variance	3	1, 2, 4	Study 3
11.	Prior degrees of freedom	2	5, 15	Study 3

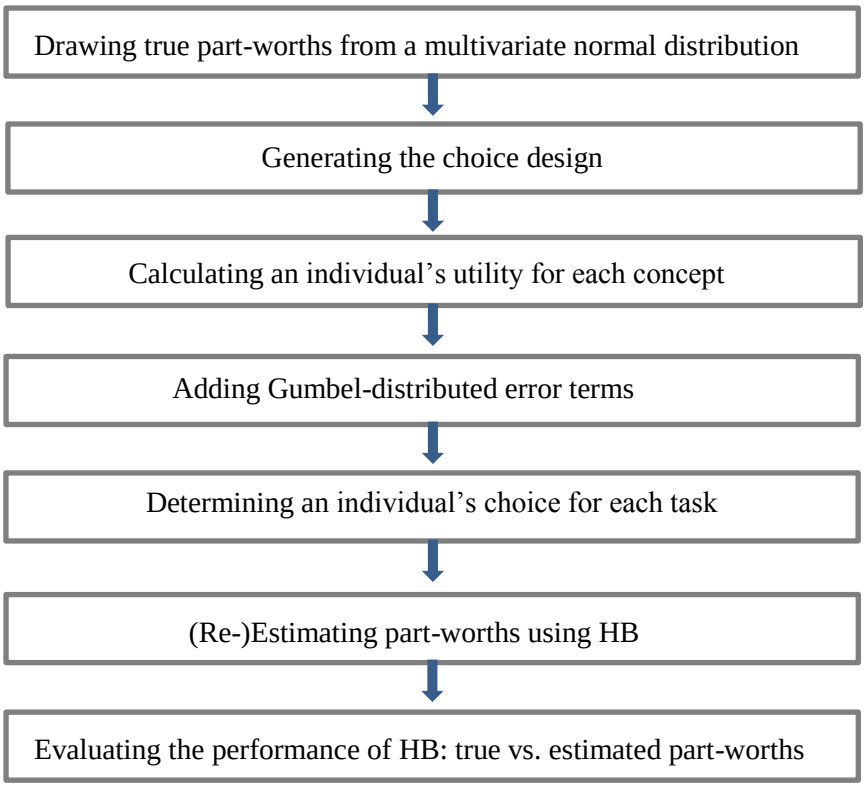
V_β is specified as a diagonal matrix and captures the amount of heterogeneity in the part-worths across individuals.[7] Specifically, the data generation process leans on Wirth (2010) and takes the following steps: First, for each treatment, the vector of population means ($\overline{\beta}$) was randomly drawn. 80% of the mean betas were randomly generated from the range between −2 to 2. To get part-worths that are somewhat more extreme, another 10% of the mean betas were randomly generated to fall in the ranges between −5 to −2 and 2 to 5, respectively. Such a distribution of mean betas is typical of that observed in empirical applications, a finding that we can confirm based on an inspection of a random sample of 250 real-world HB-CBC studies conducted at TNS Infratest (with 6 to 12 attributes, 3 to 5 attribute levels, and 11 to 15 choice tasks). Second, the covariance matrix V_β was generated to approximate a multivariate normal distribution, where the variances along the main diagonal are allowed to differ between attributes but should turn out smaller for homogeneous samples as compared to heterogeneous samples (factor 5). In particular, the main-diagonal elements for homogeneous [heterogeneous] samples were generated from a mixture of gamma and uniform draws according to the following steps: (1) Random draws (R1) were generated from a gamma distribution with shape parameter 0.7 [0.7] and scale parameter 1.5 [4.5]. (2) Since the gamma distribution is highly skewed with large parts of its mass near zero, additional random draws (R2) were generated from a uniform distribution U(0.08,0.4) [U(0.2,2)] and added to R1 in order to avoid variances (R1) that are too small on the one hand. (3) To avoid variances resulting from the sum of R1 and R2 that are too large on the other hand, further random draws (R3) were generated from a second uniform distribution U(9,11) [U(13,18)]. (4) Lastly, the minima of (R1+R2) and R3 were chosen to determine the main-diagonal elements. Overall, for H attributes with I levels each, a H times (I-1)-dimensional vector of variances is drawn from a mixture of gamma and uniform distributions.

Given the generated "true" individual-level part-worths β_n, deterministic utilities of the *n*-th respondent (factor 4) for each alternative (factor 7) in each choice task (factor 3) were computed as $U_n = X\beta_n$. Finally, a Gumbel-distributed error term (factor 6) was added to U_n in order to obtain the stochastic utilities. Based on the simulated respondents' choices the individual-level part-worths (as well as the population mean betas $\overline{\beta}$ and the covariance matrix

[7] Following previous conjoint simulation studies, we focus on main-effects models und do not include interactions between attributes. Consequently, the off-diagonal elements of the covariance matrix V_β are set to zero (e.g. compare Andrews, Ansari, and Currim 2002; Wirth 2010).

V_β) were then re-estimated using HB. Figure 3.1 displays the steps of the data generation process.

Figure 3.1: Data generation process



We used a total of 200,000 MCMC iterations (study 2: 204,000 iterations), where we chose 190,000 iterations (study 2: 199,000) for the burn-in phase and 10,000 iterations (study 2: 5,000) after convergence. Using such a large number of burn-in iterations ensures the convergence of the Markov chain to the posterior distribution. To reduce the amount of correlation across the draws only every tenth draw (study 2: every fifth draw) is retained to compute the point estimates for the individual-level parameters. Further, in order to ensure that convergence of the Markov chain to the posterior distribution has been achieved, convergence was formally tested by using the Gelman-Rubin Potential Scale Reduction Factor (Gelman and Rubin 1992; Brooks and Gelman 1998).

3.2. Measures of performance

To assess the performance of HB-CBC, we used different measures for parameter recovery, goodness-of-fit and predictive accuracy. We next describe the performance measures in more detail. As the performance measures differ from simulation study to simulation study Table 3.2 gives an overview of the specific performance measures used in each simulation study.

Parameter recovery

Parameter recovery can be measured by the mean Pearson correlation between true and re-estimated part-worths. To compute the mean correlation across respondents, the individual coefficients have to be at first rescaled using Fisher's z-transformation as Pearson correlations are not interval-scaled.

Second, parameter recovery can be measured by the root mean square error (RMSE) between the true part-worths (β) and the re-estimated part-worths ($\hat{\beta}$):

$$\text{(3.1)} \quad \text{RMSE}(\beta) = \sqrt{\frac{\sum_n \sum_h \sum_i (\hat{\beta}_{nhi} - \beta_{nhi})^2}{NHI}},$$

where N refers to the number of respondents, and H and I denote the number of attributes and the number of attribute levels, respectively. Following Andrews, Ainslie, and Currim (2002) we divided the true part-worths in the high error variance condition $(\mu = \sqrt{2})$ by the square root of two before the RMSE (β) is computed, since the scale factor is confounded with the parameter values and the logit model implicitly assumes a scale factor of one for estimation. To make the resulting RMSE values in the high error variance condition $(\mu = \sqrt{2})$ comparable to those in the standard error variance condition $(\mu = 1)$ we later multiply them by the square root of two (cf. Andrews, Ainslie, and Currim 2002).

In addition, the 95% Bayesian credible intervals for estimated parameters obtained from the draws of the posterior distribution can be determined (Greenberg 2008; Gelman et al. 2008). Accordingly, as our third measure of parameter recovery we computed the percentage of true betas (referred to as %TrueBetas) lying within the respective credible intervals across respondents.

Goodness-of-fit

Goodness-of-fit can be measured by the log-likelihood divided by the number of observations in the data (number of respondents times number of choice tasks). The correction by the number of observations enables the comparison of the log-likelihood across different treatments (cf. Andrews, Ainslie, and Currim 2002):

$$(3.2) \quad \ln(L(\hat{\beta})) = \frac{\sum_n \sum_k \sum_j Y_{nkj} \ln(\hat{P}_{nkj})}{NK},$$

where $\hat{\beta}$ is the vector of re-estimated part-worths, Y_{nkj} is a binary variable indicating whether respondent *n* has chosen alternative *j* from choice task *k* or not, $\hat{P}_{nkj}$ is the estimated choice probability for respondent *n* with respect to alternative *j* in choice task *k*, and *N* and *K* denote the number of respondents and the number of choice tasks per respondent, respectively.

Further measures of model fit are the Brier score[8] (Brier 1950; Gneiting and Raftery 2007; Roulston 2007; Kneib, Baumgartner, and Steiner 2007) and the spherical score (Gneiting and Raftery 2007; Kneib, Baumgartner, and Steiner 2007). Similar to the log-likelihood, both measures are corrected by the number of observations to enable a fair comparison across treatments with different numbers of respondents, choice tasks per respondent, and alternatives per choice tasks. Brier and spherical scores are defined as follows:

$$(3.3) \quad \text{Brier score}(\hat{\beta}) = -\frac{\sum_n \sum_k \sum_j (Y_{nkj} - \hat{P}_{nkj})^2}{NKJ},$$

$$(3.4) \quad \text{Spherical score}(\hat{\beta}) = \frac{\sum_n \sum_k \left(\hat{P}_{nkj^*} / \sqrt{\sum_{j=1}^{J} (\hat{P}_{nkj})^2} \right)}{NK},$$

where j^* denotes that alternative in choice task *k* that was actually chosen by respondent *n*.

[8] The Brier score is commonly used as scoring rule to evaluate probability forecasts in meteorology. Basically, the Brier score can be applied to any case where the estimated probability that an event will occur is compared to whether the event actually occurred or not.

Contrary to the log-likelihood, the Brier score directly compares estimated choice probabilities $\hat{P}_{nkj}$ to the actual choice pattern Y_{nkj} of respondents. Moreover, both Brier and spherical score utilize the entire predictive distribution of choice probabilities, whereas the log-likelihood only considers choice probabilities of alternatives chosen by respondents (Y_{nkj} = 1) but not of alternatives not chosen by respondents (Y_{nkj} = 0). The log-likelihood does therefore not fully exploit the information contained in the predictive distribution.

Goodness-of-fit can also be measured by the percent certainty, which is equivalent to the likelihood ratio index (McFadden 1974; Hauser 1978; Ben-Akiva and Lerman 1985):

$$(3.5) \quad \rho^2 = 1 - \frac{\ln L(\hat{\beta})}{\ln L(0)} \qquad \rho^2 \in [0,1],$$

where $\ln L(\hat{\beta})$ is the log-likelihood of the model estimated with the vector of re-estimated part-worths $\hat{\beta}$, and $\ln L(0)$ is the log-likelihood of the null model. The percent certainty measure acts like a pseudo-R^2 with a value of zero indicating that the model fits the data at only the chance level (ρ^2 = 0 when $\ln L(\hat{\beta}) = \ln L(0)$), and with a value of one indicating a perfect fit (ρ^2 = 1 when $\ln L(\hat{\beta}) = 0$), otherwise $0 < \rho^2 < 1$ (cf. Hauser 1978).

Predictive performance

According to Winkler and Murphy (1992) there exists no single best forecasting measure. It is therefore reasonable to use alternative statistics for measuring the predictive accuracy, each of them having somewhat different strengths and weaknesses. Under each treatment, we compare the predicted shares of choice ($\hat{W}_j$) based on the re-estimated part-worths to the "true" shares of choice (W_j) based on the generated part-worths across alternatives *j* in holdout task *k* along the following measures of predictive accuracy (Leeflang et al. 2000):

The *mean absolute error (MAE)* measures the average absolute deviation between true and predicted shares of choice:

$$(3.6) \quad \text{MAE (W)} = \frac{1}{J} \sum_j \left| \hat{W}_j - W_j \right|$$

By squaring the deviations, the *mean squared error (MSE)* weights large prediction errors more heavily than small prediction errors. The disproportionate influence of larger deviations is a desirable property, because larger prediction errors with regard to shares have disproportionate negative effects on managerial decisions (Chakraborty et al. 2002):

$$\text{(3.7)} \quad \text{MSE (W)} = \frac{1}{J}\sum_{j}\left(\hat{W}_j - W_j\right)^2$$

Taking the square root of the MSE yields the *root mean squared error (RMSE)*. Like the MSE, the RMSE penalizes larger prediction errors more strongly but finally reports the average prediction error in the dimension of the original measurement units. Therefore, the RMSE is directly interpretable in terms of measurement units:

$$\text{(3.8)} \quad \text{RMSE (W)} = \sqrt{\frac{\sum_{j}\left(\hat{W}_j - W_j\right)^2}{J}}$$

The *mean absolute percentage error (MAPE)* is a dimensionless measure. Similar to the MAE, absolute rather than squared prediction errors are computed. However, each absolute prediction error is expressed relative to the true choice share for alternative *j*:

$$\text{(3.9)} \quad \text{MAPE (W)} = \frac{1}{J}\sum_{j}\left|\frac{\hat{W}_j - W_j}{W_j}\right| \cdot 100\%$$

Note that the MAPE is undefined for $W_j = 0$, i.e. if the choice share of alternative *j* is zero.

The *relative absolute error (RAE)* compares the prediction error of a given model to the prediction error obtained from a naive forecasting model, where the latter defines a benchmark by using choice probabilities that would result purely by chance (as denoted by B_j):[9]

[9] For example, having three alternatives in each holdout task, the choice probability due to chance is 33.33% for each alternative, respectively.

$$\text{(3.10)}\quad \text{RAE (W)} = \frac{\sum_j \left|\hat{W}_j - W_j\right|}{\sum_j \left|B_j - W_j\right|}$$

If the RAE statistic is less than one the forecasting model outperforms the chance model, and if the RAE statistic is greater than one the forecasting model is worse than the chance model.

For all five measures of predictive accuracy values of zero indicate no prediction error, i.e. a perfect prediction of the "true" choice shares.

So far, predictive performance is measured at the aggregate respondent level. To measure the predictive accuracy at the individual level on the other hand, the hit rate can be computed, i.e. the percentage of first choice hits in holdout tasks (Vriens, Wedel, and Wilms 1996; Andrews, Ansari, and Currim 2002). Accordingly, a hit is counted if the alternative that was actually chosen by a respondent is correctly predicted. In addition, both the Brier score and the spherical score can be applied to the holdout tasks, as alternative measures to evaluate the predictive performance at the individual respondent level.

Table 3.2: Overview of performance measures used in each simulation study

	Recovery			Goodness-of-Fit				Predictive Accuracy							
								aggregate respondent level					individual respondent level		
Simulation study	Mean correlation	RMSE(betas)	%TrueBetas	Log-likelihood	Brier score	Spherical score	Percent Certainty	MAE	MSE	RMSE	MAPE	RAE	Hit rate	Brier score	Spherical score
Study 1: Limits for parameter settings in CBC analysis	x	x	x	x	x	x				x			x	x	x
Study 2: Using HB draws for improving predictions in conjoint simulations								x	x	x	x	x			
Study 3: Analyzing HB covariance matrix prior settings	x	x					x			x			x		

Chapter 4
Summary of results

In both marketing research theory and practice Hierarchical Bayes has become the most commonly used method to estimate individual part-worth utilities for choice-based conjoint models. The main focus of this thesis lies on the investigation of the statistical performance of HB-CBC. In particular, three studies based on synthetic choice-based conjoint data were conducted in order to systematically explore (1) if and where there are limits for HB estimation in CBC studies, (2) the accuracy of preference share predictions based on HB draws in comparison to other choice rules in the context of the IIA property, and (3) how HB prior settings affect the performance of HB-CBC models.

First, we designed a simulation study to systematically explore the limits of HB for choice-based conjoint analysis. The statistical performance of HB was evaluated under experimentally varying conditions using criteria for goodness-of-fit, parameter recovery and predictive accuracy. In particular, the impact of the seven experimental factors (number of attributes and attribute levels, number of choice tasks, number of alternatives per choice task, number of respondents, sample structure, and error variance) on each of the ten performance measures (cf. Table 3.2) was investigated by analyses of variance for both main and first-order interaction effects. The ANOVAs are based on 1296 observations with 1211 degrees of freedom for error (within-groups degrees of freedom). Table 4.1 summarizes the ANOVA results. About 90% of the main effects are highly significant ($p < .0001$) and about 60% of the first-order interaction effects are significant. However, due to the large sample size (N=1296) even very small differences may turn out significant, which does not mean that differences are actually (managerially) relevant. To further assess that relevance of effects, we calculate Eta squared (η^2) as a measure of effect size in ANOVA. To interpret the effect sizes we follow Cohen's (1988) guidelines, i.e. $\eta^2 = .01$ corresponds to a small effect, $\eta^2 = .06$ to a medium effect, $\eta^2 = .14$ to a large effect. The effect sizes are presented in Table 4.2. Accordingly, our results indicate that the number of attributes, the number of attribute levels, the number of alternatives per choice task and whether the sample's preference structure is more homogeneous or more heterogeneous seem to be the primary drivers for model performance. On the other hand, the number of choice tasks and the number of respondents seem to play only a secondary or even negligible role, with corresponding small effect sizes for main effects with regard to all performance measures. With regard to the interactions results show that for most of the first-

order interactions (except the interaction between the number of attributes and the number of attribute levels) the proportion of the total variance that is attributed to an effect is close to zero.

Further, the inspection of the means of the ten performance measures at the individual factor levels[10] shows that mean correlations and hit rates are hardly affected by variations of the number of choice tasks and the number of respondents and only moderately by the number of attributes. In order to reveal whether mean correlations or hit rates probably "break down" if the number of attributes is further increased or the number of choice tasks respectively the number of respondents is further reduced, we performed an additional sensitivity analysis with still more extreme level settings for those experimental factors. The results show that for simple CBC settings with a large amount of information available from respondents and relatively few parameters to be estimated HB estimation proves to be quite robust. In particular, our study provides evidence that holding other factors at convenient levels far more attributes than previously suggested in the relevant literature can be used in CBC studies. From a managerial point of view the findings further suggest that the sample size and/or the number of choice tasks per respondent could be held quite small, thereby enabling cost savings or preventing respondent fatigue and associated effects like simplification strategies of respondents in later tasks. However, results also demonstrate that for more complex CBC settings with an already high number of part-worths to be estimated but rather little information available from respondents, the HB model is starting to collapse if more than one of those factors (number of attributes, number of choice tasks, number of respondents) is set to an extreme level.

More detailed results of simulation study 1 can be obtained from the author upon request.

[10] For experimental factors with more than two levels post hoc tests were conducted to examine which of the factor level means significantly differ from each other. For post hoc tests the Bonferroni correction was used in order to control the familywise error by correcting the level of significance for each t-test such that the cumulative Type I error rate (α) across all comparisons remains at .05.

Table 4.1: P-values of main and interaction effects w.r.t. performance measures (study 1) (N = 1296; significant effects are bold; d.f. = degrees of freedom)

	Recovery			Goodness-of-Fit			Predictive Accuracy			
Source (d.f.)	RMSE(betas)	%TrueBetas	Mean correlation	Log-likelihood	Brier score	Spherical score	RMSE	Brier score	Spherical score	Hit rate
Number of attributes (3)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Number of attribute levels (2)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Number of choice tasks (2)	**<.0001**	**.027**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**.006**
Number of alternatives (2)	**<.0001**	.108	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Number of respondents (2)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	.647	**.027**	.184	.608
Sample structure (1)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Error variance (1)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**.001**	**<.0001**	**<.0001**	**<.0001**
Number of attributes x Number of attribute levels (6)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**.001**	**.002**	**.001**	**.003**
Number of attributes x Number of choice tasks (6)	**.027**	.485	.975	**<.0001**	**<.0001**	**<.0001**	.980	.983	.948	.946
Number of attributes x Number of alternatives (6)	**<.0001**	.252	**<.0001**	**<.0001**	**<.0001**	**<.0001**	.129	**.001**	.740	.760
Number of attributes x Number of respondents (6)	**<.0001**	**<.0001**	**.001**	**<.0001**	**<.0001**	**<.0001**	**.001**	.460	.455	.371
Number of attributes x Sample structure (3)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	.527	.128	.177	.211
Number of attributes x Error variance (3)	**<.0001**	.139	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**.073**	**.016**	**.003**	**.001**
Number of attribute levels x Number of choice tasks (4)	.371	.461	.848	**<.0001**	**<.0001**	**<.0001**	.108	.260	.232	.246

	Recovery			Goodness-of-Fit			Predictive Accuracy			
Source (d.f.)	RMSE(betas)	%TrueBetas	Mean correlation	Log-likelihood	Brier score	Spherical score	RMSE	Brier score	Spherical score	Hit rate
Number of attribute levels x Number of alternatives (4)	**.004**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Number of attribute levels x Number of respondents (4)	**<.0001**	**.003**	**.066**	**<.0001**	**<.0001**	**<.0001**	.624	.176	**.048**	**.051**
Number of attribute levels x Sample structure (2)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**.033**	.146	.142	.146
Number of attribute levels x Error variance (2)	**<.0001**	.874	.957	**<.0001**	**<.0001**	**<.0001**	.161	.570	.574	.472
Number of choice tasks x Number of alternatives (4)	.711	.955	**.079**	**<.0001**	.570	**<.0001**	**.088**	.142	**.008**	**.009**
Number of choice tasks x Number of respondents (4)	.245	.432	.497	1.000	.999	.998	.959	.326	.440	.444
Number of choice tasks x Sample structure (2)	**<.0001**	.354	**<.0001**	**<.0001**	**<.0001**	**<.0001**	.271	.841	.901	.970
Number of choice tasks x Error variance (2)	**<.0001**	.810	.330	**<.0001**	**<.0001**	**<.0001**	.302	.103	**.048**	**.044**
Number of alternatives x Number of respondents (4)	**.066**	.826	**.013**	**<.0001**	**.058**	**<.0001**	.121	**.065**	.121	**.086**
Number of alternatives x Sample structure (2)	**<.0001**	**.001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	.467	.986	.231	.233
Number of alternatives x Error variance (2)	**<.0001**	.839	.697	**<.0001**	**<.0001**	**<.0001**	.558	.892	.139	.133
Number of respondents x Sample structure (2)	**<.0001**	**<.0001**	.652	**<.0001**	**<.0001**	**<.0001**	**.019**	.195	.264	.414
Number of respondents x Error variance (2)	**<.0001**	.116	.419	**<.0001**	**<.0001**	**<.0001**	**.019**	.307	.531	.619
Sample structure x Error variance (1)	**.004**	**.031**	**.091**	**<.0001**	**<.0001**	**<.0001**	.310	.129	.108	.091
R^2 (Adjusted R^2)	.956 (.953)	.772 (.757)	.940 (.936)	.962 (.959)	.963 (.960)	.956 (.953)	.318 (.270)	.600 (.572)	.437 (.397)	.376 (.333)

Table 4.2: Effect sizes of main and interaction effects according to Cohen's guidelines (study 1)

	Recovery			Goodness-of-Fit			Predictive Accuracy			
Source (d.f.)	RMSE (betas)	%TrueBetas	Mean correlation	Log-likelihood	Brier score	Spherical score	RMSE	Brier score	Spherical score	Hit rate
Number of attributes (3)	large	large	medium	large	large	large	small	medium	medium	medium
Number of attribute levels (2)	large	small	large	large	large	large	medium	medium	medium	medium
Number of choice tasks (2)	small	small	small	small	small	small	small	small	small	small
Number of alternatives (2)	small	small	medium	medium	small	medium	small	large	small	medium
Number of respondents (2)	small	small	small	small	small	small	small	small	small	.small
Sample structure (1)	medium	large	large	small	small	small	small	small	small	small
Error variance (1)	medium	small	small	medium	medium	small	small	small	small	small
Number of attributes x Number of attribute levels (6)	small	small	small	medium	medium	medium	small	small	small	small
Number of attributes x Number of choice tasks (6)	small	small	small	small	small	small	small	small	small	small
Number of attributes x Number of alternatives (6)	small	small	small	small	small	small	small	small	small	small
Number of attributes x Number of respondents (6)	small	small	small	small	small	small	small	small	small	small
Number of attributes x Sample structure (3)	small	small	small	small	small	small	small	small	small	small
Number of attributes x Error variance (3)	small	small	small	small	small	small	small	small	small	small
Number of attribute levels x Number of choice tasks (4)	small	small	small	small	small	small	small	small	small	small
Number of attribute levels x Number of alternatives (4)	small	small	small	small	small	small	small	small	small	small
Number of attribute levels x Number of respondents (4)	small	small	small	small	small	small	small	small	small	small
Number of attribute levels x Sample structure (2)	small	small	small	small	small	small	small	small	small	small

	Recovery			Goodness-of-Fit			Predictive Accuracy			
Source (d.f.)	RMSE (betas)	%TrueBetas	Mean correlation	Log-likelihood	Brier score	Spherical score	RMSE	Brier score	Spherical score	Hit rate
Number of attribute levels x Error variance (2)	small	small	small	small	small	small	small	small	small	small
Number of choice tasks x Number of alternatives (4)	small	small	small	small	small	small	small	small	small	small
Number of choice tasks x Number of respondents (4)	small	small	small	small	small	small	small	small	small	small
Number of choice tasks x Sample structure (2)	small	small	small	small	small	small	small	small	small	small
Number of choice tasks x Error variance (2)	small	small	small	small	small	small	small	small	small	small
Number of alternatives x Number of respondents (4)	small	small	small	small	small	small	small	small	small	small
Number of alternatives x Sample structure (2)	small	small	small	small	small	small	small	small	small	small
Number of alternatives x Error variance (2)	small	small	small	small	small	small	small	small	small	small
Number of respondents x Sample structure (2)	small	small	small	small	small	small	small	small	small	small
Number of respondents x Error variance (2)	small	small	small	small	small	small	small	small	small	small
Sample structure x Error variance (1)	small	small	small	small	small	small	small	small	small	small

In our second simulation study, we focused on the application of HB-CBC for predicting consumer preferences. We designed a simulation study to systematically compare preference share predictions based on the following choice rules: (1) the First Choice rule, (2) the Logit Choice rule, (3) the Randomized First Choice rule, and (4) HB random draws combined with first choice simulations. In particular, the study wants to shed more light on the question which choice rule should be used in order to predict preference shares as accurate as possible. Further, to assess how well the choice rules tolerate the IIA property, two different holdout choice scenarios were designed. Once, holdout tasks were designed to have two of the three alternatives extremely similar to each other so that predictions are highly prone to the IIA property. Alternatively, holdout tasks were designed to contain alternatives that are all different from each other in their profiles. Thus, in total ten ANOVAs (5 measures of predictive accuracy times 2 holdouts) were conducted, each based on 288 observations with 263 degrees of freedom for error. Since the ANOVA homogeneity of variance assumption was oftentimes not met, we applied the Box correction for heterogeneous variances to ensure that this violation does not really affect the F-statistic seriously. The Box approximation corrects the F-statistic by adjusting the degrees of freedom (for more details, see Box 1954). As can be seen from Table 4.3 and Table 4.4, the F-Test is highly robust because there are not any substantial differences with regard to the significance of main and first-order interactions effects. The results of ANOVAs can be summarized as follows: All main effects turn out highly significant independent whether two very similar alternatives are contained in the holdout choice scenario (Table 4.3, type SIMILAR) or not (Table 4.4, type DISSIMILAR). With regard to the first-order interactions results show significant effects between the factors sample structure and model complexity and between the factors sample structure and number of choice tasks for both types of holdout choice scenarios and nearly across all predictive measures. The corresponding effect sizes of the main and first-order interaction effects are reported in Table 4.5 and Table 4.6. Overall, the three experimental factors choice rule, number of choice tasks, and model complexity turn out to be primary drivers for predictive performance. Further, the means of the five measures of predictive accuracy depending on the experimental condition (i.e. for each factor level) were examined. For experimental factors with more than two levels post hoc tests were conducted by using the Bonferroni correction (see footnote 10). The comparison between the two types of holdout choice scenarios revealed that prediction errors are higher when two very similar alternatives exist, because then predictions are more prone to the IIA property. The major finding of the study is that using HB draws for first choice simulations leads to the lowest predictions errors across all choice rules independent of the type of

holdout choice scenario considered. Therefore, when HB is used for part-worth estimation, HB draws should be saved and directly employed for preference simulations.

More detailed results of simulation study 2 can be obtained from the author upon request.

Table 4.3: P-values of main and interaction effects on performance measures w.r.t. holdout choice scenarios where two out of three alternatives are extremely SIMILAR (study 2) (N = 288; p-values of the Box correction are in bold and in parentheses; d.f. = degrees of freedom)

	Predictive Accuracy				
Source (d.f.)	MAE	MSE	RMSE	MAPE	RAE
Choice rule (3)	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**
Number of choice tasks (2)	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**
Sample structure (1)	.002 **(.002)**	.005 **(.005)**	.002 **(.002)**	<.0001 **(<.0001)**	.002 **(.002)**
Model complexity (1)	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**
Sample structure x Model complexity (1)	.006 **(.008)**	.183 **(.182)**	.003 **(.004)**	.006 **(.008)**	.006 **(.008)**
Sample structure x Choice rule (3)	.498 **(.486)**	.419 **(.397)**	.519 **(.509)**	.250 **(.252)**	.496 **(.485)**
Sample structure x Number of choice tasks (2)	.059 **(.064)**	.028 **(.044)**	.088 **(.093)**	.025 **(.029)**	.059 **(.064)**
Model complexity x Choice rule (3)	.264 **(.266)**	.159 **(.180)**	.344 **(.339)**	.299 **(.298)**	.262 **(.264)**
Model complexity x Number of choice tasks (2)	.346 **(.338)**	.079 **(.102)**	.431 **(.418)**	.606 **(.580)**	.346 **(.339)**
Choice rule x Number of choice tasks (6)	.284 **(.289)**	.146 **(.180)**	.303 **(.306)**	.387 **(.385)**	.284 **(.289)**
R^2 (Adjusted R^2)	.496 (.449)	.429 (.377)	.502 (.456)	.520 (.477)	.496 (.450)

Table 4.4: P-values of main and interaction effects on performance measures w.r.t. holdout choice scenarios with DISSIMILAR alternatives (study 2) (N = 288; p-values of the Box correction are in bold and in parentheses; d.f. = degrees of freedom)

	Predictive Accuracy				
Source (d.f.)	MAE	MSE	RMSE	MAPE	RAE
Choice rule (3)	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**
Number of choice tasks (2)	<.0001 **(<.0001)**	<.0001) **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**
Sample structure (1)	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**
Model complexity (1)	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**	<.0001 **(<.0001)**
Sample structure x Model complexity (1)	<.0001 **(<.0001)**	.007 **(.010)**	<.0001 **(<.0001)**	.049 **(.051)**	<.0001 **(<.0001)**
Sample structure x Choice rule (3)	.620) **(.600)**	.425 **(.408)**	.635 **(.616)**	.031 **(.033)**	.620 **(.600)**
Sample structure x Number of choice tasks (2)	<.0001 **(<.0001)**	.001 **(.002)**	<.0001 **(<.0001)**	.013 **(.014)**	<.0001 **(<.0001)**
Model complexity x Choice rule (3)	.392 **(.388)**	.204) **(.212)**	.344 **(.342)**	.114 **(.118)**	.393 **(.389)**
Model complexity x Number of choice tasks (2)	.788 **(.746)**	.242 **(.241)**	.696 **(.656)**	<.0001 **(.001)**	.789 **(.746)**
Choice rule x Number of choice tasks (6)	.339 **(.340)**	.194 **(.219)**	.313 **(.315)**	.497 **(.491)**	.341 **(.341)**
R^2 (Adjusted R^2)	.498 (.453)	.413 (.359)	.507 (.462)	.605 (.569)	.498 (.452)

Table 4.5: Effect sizes of main and interaction effects for holdout choice scenarios where two out of three alternatives are extremely SIMILAR according to Cohen's guidelines (study 2)

	Predictive Accuracy				
Source (d.f.)	MAE	MSE	RMSE	MAPE	RAE
Choice rule (3)	large	medium	large	large	large
Number of choice tasks (2)	medium	medium	large	large	medium
Sample structure (1)	small	small	small	small	small
Model complexity (1)	medium	medium	medium	medium	medium
Sample structure x Model complexity (1)	small	small	small	small	small
Sample structure x Choice rule (3)	small	small	small	small	small
Sample structure x Number of choice tasks (2)	small	small	small	small	small
Model complexity x Choice rule (3)	small	small	small	small	small
Model complexity x Number of choice tasks (2)	small	small	small	small	small
Choice rule x Number of choice tasks (6)	small	small	small	small	small

Table 4.6: Effect sizes of main and interaction effects for holdout choice scenarios with DISSIMILAR alternatives according to Cohen's guidelines (study 2)

	Predictive Accuracy				
Source (d.f.)	MAE	MSE	RMSE	MAPE	RAE
Choice rule (3)	large	medium	large	large	large
Number of choice tasks (2)	medium	medium	medium	medium	medium
Sample structure (1)	small	small	small	medium	small
Model complexity (1)	medium	medium	medium	medium	medium
Sample structure x Model complexity (1)	small	small	small	small	small
Sample structure x Choice rule (3)	small	small	small	small	small
Sample structure x Number of choice tasks (2)	small	small	small	small	small
Model complexity x Choice rule (3)	small	small	small	small	small
Model complexity x Number of choice tasks (2)	small	small	small	small	small
Choice rule x Number of choice tasks (6)	small	small	small	small	small

So far, default prior parameter settings were used for HB estimation. However, the choice of the HB prior parameter settings (the prior variance and the prior degrees of freedom) may exert an influence on Bayesian estimates depending on the characteristics of the respective data set. In simulation study 3, we therefore investigated the effects of both the prior degrees of freedom and the prior variance settings on the posterior Bayesian estimates, as well as the interaction of these two prior settings based on simulated CBC data. We experimentally manipulated seven factors (the prior variance, the prior degrees of freedom, number of respondents, number of attributes, number of choice tasks, sample structure, as well as the error variance) and assessed the statistical performance of HB-CBC using criteria for parameter recovery, goodness-of-fit and predictive accuracy. Similar to the first and second simulation study the impact of the experimental factors on each performance measure was investigated by analyses of variance for both main effects and first-order interaction effects. The ANOVAs are based on 432 observations with 379 degrees of freedom for error. Table 4.7 summarizes the ANOVA results. About 74% of the main effects are highly significant beyond .0001, another 14% are still significant at $p < .05$, and only 3 main effects are not significant ($p > .10$). Further, about 42% of the first-order interaction effects between factors turn out significant at $p < .05$, while 53% of the interaction effects are not significant (see Table 4.7). With regard to the main effects both the prior variance ($p = .412$) and the prior degrees of freedom ($p = .945$) do not show a significant impact on the accuracy of shares of choice predictions measured by the RMSE. In addition, the results of the ANOVAs show no statistically significant interaction effect between the prior variance and the prior degrees of freedom on shares of choice predictions. Further, based on Cohen's eta squared only very small interaction effects were observed between the prior variance and the remaining experimental factors across the performance measures, with corresponding effect sizes for interactions being approximately zero (despite some interaction effects are significant, compare Table 4.7 and Table 4.8). The same applies to the interaction effects concerning the prior degrees of freedom. Therefore, results indicate that interaction effects related to the two prior settings play only a secondary if not a negligible role. As can also be seen from Table 4.8 noticeable effect sizes are observed for only some of the experimental factors. Results show that the number of attributes, the number of respondents, the error variance, the sample structure and the prior variance turned out to be primary drivers for model performance. On the other hand, the number of choice tasks, the error variance, and importantly the prior degrees of freedom seem to play only secondary if not negligible roles as there is not one large effect size across all five performance measures for these three factors. Thus, one major finding is that the prior degrees of freedom do not

have a noticeable impact on the performance of HB. The inspection of the factor level means for the prior variance shows that with respect to parameter recovery and model fit HB tends to slightly overfit the data for a prior variance of 4 while there is not much difference in the performance of HB for a prior variance of 1 or 2. However, despite that overfitting for a prior variance of 4 the predictive accuracy of HB-CBC does not suffer. Therefore, we further performed a sensitivity analysis with extreme level settings for the prior variance (we increase the prior variance settings stepwise to 10, 20 and 50). The most striking finding of the sensitivity analysis is that, although overfitting problems become obvious, the predictive performance of HB-CBC is again not markedly affected by an increase of the prior variance up to high values, such as 50.

More detailed results of simulation study 3 can be obtained from the author upon request.

Table 4.7: P-values of main and interaction effects on performance measures (study 3) (N = 432; significant effects are bold; d.f. = degrees of freedom)

	Recovery		Goodness-of-Fit	Predictive Accuracy	
Source (d.f.)	Mean correlation	RMSE (betas)	Percent certainty	RMSE	Hit rate
Prior variance (2)	**.011**	**<.0001**	**<.0001**	.412	**.012**
Prior degrees of freedom (1)	**<.0001**	**<.0001**	**<.0001**	.945	**.008**
Number of respondents (2)	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Number of attributes (2)	**<.0001**	**<.0001**	**<.0001**	**.007**	**<.0001**
Number of choice tasks (1)	**<.0001**	**<.0001**	**<.0001**	**.097**	**<.0001**
Sample structure (1)	**<.0001**	**<.0001**	**<.0001**	**.001**	**<.0001**
Error variance (1)	**<.0001**	**<.0001**	**<.0001**	.918	**<.0001**
Prior variance x Prior degrees of freedom (2)	.378	**<.0001**	**<.0001**	.247	.148
Prior variance x Number of respondents (4)	.447	**<.0001**	**<.0001**	.314	.540
Prior variance x Number of attributes (4)	.110	**<.0001**	**<.0001**	**.099**	.511
Prior variance x Number of choice tasks (2)	.568	.141	**.004**	.677	.607
Prior variance x Sample structure (2)	.761	**<.0001**	**<.0001**	.160	.745
Prior variance x Error variance (2)	.444	**<.0001**	**<.0001**	.366	.989
Prior degrees of freedom x Number of respondents (2)	**.012**	**<.0001**	**<.0001**	.361	.187
Prior degrees of freedom x Number of attributes (2)	**<.0001**	**<.0001**	.143	.460	.658
Prior degrees of freedom x Number of choice tasks (1)	.585	**<.0001**	**.001**	.925	.441
Prior degrees of freedom x Sample structure (1)	.225	**<.0001**	**.008**	.442	.884

	Recovery		Goodness-of-Fit	Predictive Accuracy	
Source (d.f.)	Mean correlation	RMSE (betas)	Percent certainty	RMSE	Hit rate
Prior degrees of freedom x Error variance (1)	.819	**<.0001**	.740	.713	.339
Number of respondents x Number of attributes (4)	**.004**	**<.0001**	**<.0001**	.815	.608
Number of respondents x Number of choice tasks (2)	.709	**.015**	.962	.237	.738
Number of respondents x Sample structure (2)	**<.0001**	**<.0001**	**.012**	.136	.440
Number of respondents x Error variance (2)	.565	**<.0001**	**.006**	.969	.518
Number of attributes x Number of choice tasks (2)	.117	**.059**	**<.0001**	.139	.375
Number of attributes x Sample structure (2)	**<.0001**	**<.0001**	.518	**.090**	**<.0001**
Number of attributes x Error variance (2)	**.023**	**<.0001**	**<.0001**	.309	**<.0001**
Number of choice tasks x Sample structure (1)	**<.0001**	.146	**.089**	.276	**<.0001**
Number of choice tasks x Error variance (1)	.279	**<.0001**	**<.0001**	.702	.326
Sample structure x Error variance (1)	.754	**<.0001**	**<.0001**	.781	**.077**
R^2 (Adjusted R^2)	.960 (.954)	.978 (.975)	.976 (.973)	.737 (.701)	.865 (.846)

Table 4.8: Effect sizes of main and interaction effects according to Cohen's guidelines (study 3)

	Recovery		Goodness-of-Fit	Predictive Accuracy	
Source (d.f.)	Mean correla-tion	RMSE(betas)	Percent certainty	RMSE	Hit rate
Prior variance (2)	small	large	large	small	small
Prior degrees of freedom (1)	small	small	small	small	small
Number of respondents (2)	medium	medium	small	large	small
Number of attributes (2)	medium	large	.424	small	large
Number of choice tasks (1)	small	small	small	small	medium
Sample structure (1)	large	medium	small	small	large
Error variance (1)	small	medium	small	small	small
Prior variance x Prior degrees of freedom (2)	small	small	small	small	small
Prior variance x Number of respondents (4)	small	small	small	small	small
Prior variance x Number of attributes (4)	small	small	small	small	small
Prior variance x Number of choice tasks (2)	small	small	small	small	small
Prior variance x Sample structure (2)	small	small	small	small	small
Prior variance x Error variance (2)	small	small	small	small	small
Prior degrees of freedom x Number of respondents (2)	small	small	small	small	small
Prior degrees of freedom x Number of attributes (2)	small	small	small	small	small
Prior degrees of freedom x Number of choice tasks (1)	small	small	small	small	small
Prior degrees of freedom x Sample structure (1)	small	small	small	small	small
Prior degrees of freedom x Error variance (1)	small	small	small	small	small

	Recovery		Goodness-of-Fit	Predictive Accuracy	
Source (d.f.)	Mean correla-tion	RMSE(betas)	Percent certainty	RMSE	Hit rate
Number of respondents x Number of attributes (4)	small	small	small	small	small
Number of respondents x Number of choice tasks (2)	small	small	small	small	small
Number of respondents x Sample structure (2)	small	small	small	small	small
Number of respondents x Error variance (2)	small	small	small	small	small
Number of attributes x Number of choice tasks (2)	small	small	small	small	small
Number of attributes x Sample structure (2)	small	small	small	small	small
Number of attributes x Error variance (2)	small	small	small	small	small
Number of choice tasks x Sample structure (1)	small	small	small	small	small
Number of choice tasks x Error variance (1)	small	small	small	small	small
Sample structure x Error variance (1)	small	small	small	small	small

Chapter 5
Limitations and outlook

The objective of this section is to finally summarize general limitations of all three simulation studies as well as the particular limitations of each simulation study.

The benefit of a simulation study with synthetic data is that estimation results can be contrasted with the true set of parameters. That way, we were able to evaluate the performance of HB-CBC under different experimental conditions. However, real-world choice data often do not strictly adhere to the assumptions by which data are generated in simulation studies. For example the data generation process of all three simulation studies is consistent with the assumption of normally distributed preferences as supposed in the standard HB approach. Thus, the question arises how the performance of HB-CBC will be affected when this assumption is violated, especially in the context of simulation study 1. For example, respondent's preferences may be distributed according to gamma and mixtures of normal distributions (as e.g. in Andrews, Ainslie, and Currim 2002, or in Andrews, Ansari, and Currim 2002). Another aspect with regard to the data generation process that would have an impact on the obtained results is the specification of the range of the true part-worths as well as the determination of the amount of heterogeneity in the data. For example, it may depend on the magnitude of the generated part-worths how much influence the stochastic error term would have on the simulated choices. In order to get a distribution of mean betas that is typical of that observed in empirical applications, in our simulation studies we randomly generated most of the mean betas from the range between −2 to 2, another minor percentage of the mean betas was generated to fall in wider ranges to get part-worths that are somewhat more extreme. When the true part-worths are drawn from uniform distributions with smaller ranges (e.g. Andrews, Ainslie, and Currim 2002; Andrews, Ansari, and Currim 2002; Chakraborty et al. 2002), utilities would be closer together resulting in a higher impact of the stochastic component which in turn may affect measures of performance. Further, due to complexity reasons we considered only a limited number of experimental factors in our simulation studies. While the error variance represents a factor in the simulation design (either varied as in simulation study 1 and 3 or held constant as in simulation study 2) to address that the choice behavior of respondents may be more or less stochastic, behavioral effects such as respondent fatigue and associated effects like simplification strategies of respondents were not taken into account. In practice,

respondents may learn how to better complete choice tasks or may become disengaged as the number of choice tasks increases resulting in less reliable data (Kurz and Binner 2012). Therefore, additional behavioral effects such as respondent fatigue or simplification strategies of respondents that more directly can mimic real-world choice behavior could be considered for a simulation study. In contrast to the global stochastic error term that operates more or less evenly across alternatives and choice sets, those behavioral effects primarily occur in later choice tasks. Moreover, not all experimental factors that are varied in our simulation studies can be controlled by the researcher. Design factors that can be manipulated by the researcher are, for example, the sample size, the number of attributes and the number of choice tasks. However, respondent heterogeneity as well as the error variance cannot be controlled in real-world CBC studies.

With regard to the first simulation study another aspect not considered in the simulation design is the inclusion of interactions between attributes, such as those between brand and price. The inclusion of interaction terms would lead to an increase of the number of parameters which in turn may have a negative effect on the performance of the HB-CBC model. The second study focuses on the prediction of shares of preference and clearly shows the superiority of HB draws. Another issue in the field of choice rules comparisons would be to investigate the predictive performance of the choice rules in the absence of HB draws, i.e. in the case of aggregate logit estimates or when estimates from traditional latent class analysis are available. In the context of aggregate models, the RFC rule might provide greater benefits as pointed out by Huber, Orme, and Miller (1999). Moreover, empirical data could be used to assess the performance of the choice rules depending on the product category (e.g. as suggested in Arenoe 2003). For instance, predictive accuracy of the First Choice rule might be better for non-routine purchases (durables like automobiles or personal computers). Contrary, in case of frequently purchased consumer goods (e.g., beverages), respondents' choice behavior may be more probabilistic so that preferences vary over use occasions (e.g. Green and Krieger 1988; Elrod and Kumar 1989; Rao 2014). Moreover, we did not consider other approaches for addressing IIA troubles like the Nested Logit Model or the Multinomial Probit Model. Finally, the third study addresses the effects of both the prior degrees of freedom and the prior variance settings on posterior Bayesian estimates. Since results do not show a noticeable impact of the prior degrees of freedom on the performance of HB, it would be interesting from a statistical point of view to examine for still higher levels of the prior degrees of freedom as to when there would be an impact. Furthermore, measures of performance indicate that HB tends to overfit the data for an increasing prior variance. Additionally, we would be able to reveal

overfitting by using the aggregate MNL model for part-worth estimation as a benchmark. There would be strong evidence that HB overfits the data when HB performed poorly in predicting choice behavior compared to the aggregate MNL model (e.g. Pinnell and Fridley 2001).

Bibliography

Allenby, Greg M., Neeraj Arora, and James L. Ginter (1995), "Incorporating Prior Knowledge into the Analysis of Conjoint Studies," *Journal of Marketing Research*, 32 (2), 152–162.

Allenby, Greg M., Geraldine Fennel, Joel Huber, Thomas Eagle, Timothy J. Gilbride, Dan Horsky, Jaehwan Kim, Peter Lenk, Rich Johnson, Elie Ofek, Bryan Orme, Thomas Otter, and Joan Walker (2005), "Adjusting Choice Models to Better Predict Market Behavior," *Marketing Letters*, 16 (3-4), 197–208.

Allenby, Greg M. and James L. Ginter (1995), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32 (4), 392–403.

Allenby, Greg M. and Peter E. Rossi (2006), "Hierarchical Bayes Models," in *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*, Rajiv Grover and Marco Vriens, eds. Thousand Oaks, CA: Sage Publications, 418–440.

Andrews, Rick L., Andrew Ainslie, and Imran S. Currim (2002), "An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity," *Journal of Marketing Research*, 39 (4), 479–487.

Andrews, Rick L., Asim Ansari, and Imran S. Currim (2002), "Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery," *Journal of Marketing Research*, 39 (1), 87–98.

Arenoe, Bjorn (2003), "Determinants of External Validity in CBC," in *Sawtooth Software Conference Proceedings*. Sequim, WA: Sawtooth Software Inc., 217–232.

Backhaus, Klaus, Thomas Hillig, and Robert Wilken (2007), "Predicting Purchase Decisions with Different Conjoint Analysis Methods. A Monte Carlo Simulation," *International Journal of Market Research*, 49 (3), 341–364.

Baier, Daniel and Wolfgang Gaul (2007), "Market Simulation Using a Probabilistic Ideal Vector Model for Conjoint Data," in *Conjoint Measurement: Methods and Applications*, 4th edition, Anders Gustafsson, Andreas Herrmann, and Frank Huber, eds. Berlin: Springer, 47–65.

Baier, Daniel and Wolfgang Polasek (2003), "Market Simulation Using Bayesian Procedures in Conjoint Analysis," in *Exploratory Data Analysis in Empirical Research*, Manfred Schwaiger and Otto Opitz, eds. Berlin: Springer, 413–421.

Ben-Akiva, Moshe E. and Steven R. Lerman (1985): *Discrete Choice Analysis. Theory and Application to Travel Demand*. Cambridge, MA: MIT Press (MIT Press series in transportation studies, 9).

Box, George E.P. (1954), "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification," *Annals of Mathematical Statistics*, 25 (2), 290–302.

Brier, Glenn W. (1950), "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78 (1), 1–3.

Brooks, Stephan P. and Andrew Gelman (1998), "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, 7 (4), 434–455.

Chakraborty, Goutam, Dwayne Ball, Gary J. Gaeth, and Sunkyu Jun (2002), "The Ability of Ratings and Choice Conjoint to Predict Market Shares. A Monte Carlo Simulation," *Journal of Business Research*, 55 (3), 237–249.

Cohen, Jacob (1988), *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, Steven H. (1997), "Perfect Union: CBCA Marries the Best of Conjoint and Discrete Choice Models," *Marketing Research*, 9 (1), 12–17.

Elrod, Terry and S. Krishna Kumar (1989), "Bias in the First Choice Rule for Predicting Share," in *Sawtooth Software Conference Proceedings*. Ketchum, ID: Sawtooth Software Inc., 259–271.

Finkbeiner, Carl T. (1988), "Comparison of Conjoint Choice Simulators," in *Sawtooth Software Conference Proceedings*. Ketchum, ID: Sawtooth Software Inc., 75–103.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2008), *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, Andrew and Donald B. Rubin (1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7 (4), 457–511.

Gneiting, Tilmann and Adrian E. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102 (477), 359–378.

Green, Paul E. and Abba M. Krieger (1988), "Choice Rules and Sensitivity Analysis in Choice Simulators," *Journal of the Academy of Marketing Science*, 16 (1), 114–127.

Green, Paul E., Abba M. Krieger, and Yoram J. Wind (2001), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Interfaces: An International Journal of the Institute for Operations Research*, 31 (2), 56–73.

Green, Paul E. and Vithala R. Rao (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research*, 8 (3), 355–363.

Green, Paul E. and Venkat Srinivasan (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (2), 103–123.

Greenberg, Edward (2008), *Introduction to Bayesian Econometrics*. New York: Cambridge University Press.

Hauser, John R. (1978), "Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach," *Operations Research*, 26 (3), 406–421.

Huber, Joel, Bryan Orme, and Richard Miller (1999), "Dealing with Product Similarity in Conjoint Simulations," *Sawtooth Software Research Paper Series*, Sequim, WA: Sawtooth Software, Inc.

Huber, Joel and Kenneth Train (2001), "On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths," *Marketing Letters*, 12 (3), 259–269.

Kneib, Thomas, Bernhard Baumgartner, and Winfried J. Steiner (2007), "Semiparametric Multinomial Logit Models for Analyzing Consumer Choice Behaviour," *Advances in Statistical Analysis*, 91 (3), 225–244.

Kurz, Peter and Stefan Binner (2012), "'The Individual Choice Task Threshold' Need for Variable Number of Choice Tasks," in *Sawtooth Software Conference Proceedings*. Orlando, FL: Sawtooth Software, Inc., 111–127.

Leeflang, Peter S.H., Dick R. Wittink, Michael Wedel, and Philippe A. Naert (2000), *Building Models for Marketing Decisions*. Boston: Kluwer Academic Publishers.

Lenk, Peter J., Wayne S. Desarbo, Paul E.Green, and Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2), 173–191.

Lenk, Peter J. and Bryan Orme (2009), "The Value of Informative Priors in Bayesian Inference with Sparse Data," *Journal of Marketing Research*, 46 (6), 832–845.

Louviere, Jordan J. (1988), "Conjoint Analysis Modelling of Stated Preferences: A Review of Theory, Methods, Recent Developments and External Validity," *Journal of Transport Economics and Policy*, 22 (1), 93–119.

Louviere, Jordan J., Terry N. Flynn, and Richard T. Carson (2010), "Discrete Choice Experiments Are Not Conjoint Analysis," *Journal of Choice Modelling*, 3 (3), 57–72.

Louviere, Jordan J. and George Woodworth (1983), "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20 (4), 350–367.

McCullough, Paul Richard (2009), "Comparing Hierarchical Bayes and Latent Class Choice: Practical Issues for Sparse Data Sets," in *Sawtooth Software Conference Proceedings*, Delray Beach, FL: Sawtooth Software, Inc., 273–284.

McFadden, Daniel (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, Paul Zarembka, ed. New York: Academic Press, 105–142.

Orme, Bryan (2003), "New Advances Shed Light on HB Anomalies," *Sawtooth Software Research Paper Series*, Sequim, WA: Sawtooth Software, Inc.

Orme, Bryan and Gary Baker (2000), "Comparing Hierarchical Bayes Draws and Randomized First Choice for Conjoint Simulations," *Sawtooth Software Research Paper Series*, Sequim, WA: Sawtooth Software, Inc.

Orme, Bryan and Joel Huber (2000), "Improving the Value of Conjoint Simulations," *Marketing Research*, 12 (4), 12–20.

Orme, Bryan and Walter Williams (2016), "What are the Optimal HB Priors Settings for CBC and MaxDiff Studies?," *Sawtooth Software Research Paper Series*, Orem, Utah: Sawtooth Software, Inc.

Pinnell, Jon and Lisa Fridley (2001), "The Effects of Disaggregation with Partial Profile Choice Experiments," in *Sawtooth Software Conference Proceedings*. Sequim, WA: Sawtooth Software, Inc., 151–165.

R Core Team (2012), *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing (http://www.R-project.org/).

Rao, Vithala R. (2014), *Applied Conjoint Analysis*, New York, NY: Springer.

Rossi, Peter E. and Greg M. Allenby (1993), "A Bayesian Approach to Estimating Household Parameters," *Journal of Marketing Research*, 30 (2), 171–182.

Rossi, Peter E. and Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22 (3), 304–328.

Rossi, Peter E., Greg M. Allenby, and Robert E. McCulloch (2005), *Bayesian Statistics and Marketing*. Hoboken, NJ: Wiley (Wiley series in probability and statistics).

Roulston, Mark S. (2007), "Performance Targets and the Brier Score," *Meteorological Applications*, 14 (2), 185–94.

Thurstone, Louis L. (1927), "A Law of Comparative Judgement," *Psychological Review*, 34 (4), 273–286.

Train, Kenneth E. (2003), *Discrete Choice Methods with Simulation*. Cambridge, MA: Cambridge University Press.

Tsafarakis, Stelios, Evangelos Grigoroudis, and Nikolaos Matsatsinis (2011), "Consumer Choice Behaviour and New Product Development: An Integrated Market Simulation Approach," *The Journal of the Operational Research Society*, 62 (7), 1253–1267.

Vriens, Marco, Michel Wedel, and Tom Wilms (1996), "Metric Conjoint Segmentation Methods: A Monte Carlo Comparison," *Journal of Marketing Research*, 33 (1), 73–85.

Winkler, Robert L. and Allan H. Murphy (1992), "On Seeking a Best Performance Measure or a Best Forecasting Method," *International Journal of Forecasting*, 8 (1), 104–107.

Wirth, Ralf (2010), "HB-CBC, HB-Best-Worst-CBC or No HB At All?," in *Sawtooth Software Conference Proceedings*. Newport Beach, CA: Sawtooth Software, Inc., 321–356.